FAST CAMERA SELF-CALIBRATION FOR SYNTHESIZING FREE VIEWPOINT SOCCER VIDEO

Qiang Yao[†], *Akira Kubota*[††], *Kaoru Kawakita*[⋆], *Keisuke Nonaka*[†], *Hiroshi Sankoh*[†], *Sei Naito*[†]

[†] KDDI R&D Laboratories, Inc., Fujimino, Saitama, Japan
[††] Faculty of Science and Engineering, Chuo University, Tokyo, Japan
[⋆] Robit Inc., Tokyo, Japan

ABSTRACT

Recently, non-fixed camera-based free viewpoint sports video synthesis has become very popular. Camera calibration is an indispensable step in free viewpoint video synthesis, and the calibration has to be done frame by frame for a non-fixed camera. Thus, calibration speed is of great significance in real-time application. In this paper, a fast self-calibration method for a non-fixed camera is proposed to estimate the homography matrix between a camera image and a soccer field model. As far as we know, it is the first time to propose constructing feature vectors by analyzing crossing points of field lines in both camera image and field model. Therefore, different from previous methods that evaluate all the possible homography matrices and select the best one, our proposed method only evaluates a small number of homography matrices based on the matching result of the constructed feature vectors. Experimental results show that the proposed method is much faster than other methods with only a slight loss of calibration accuracy that is negligible in final synthesized videos.

Index Terms— Camera Self-Calibration, Free Viewpoint Sports Video, Homography Matrix, Field Model

1. INTRODUCTION

The appearance of Free Viewpoint Television (FTV) [1] and Free Viewpoint Video (FVV) [2] has gained increasing attention in multimedia signal processing and computer vision. [3, 4, 5, 6, 7] In the applications of FTV and FVV, it is assumed that the virtual viewpoints can be selected freely and moved around, back and forth as well as up and down, [8] bringing users an immersive and ultra-realistic experience. In addition to enhancing user experience, other applications, such as the Hawk-eye system [9] for ball tracking in the World Cup 2014 and the recently held 2016 Rio-Olympics Games, have demonstrated that free viewpoint techniques could be of great assistance in sports games and athletic competitions.

As for general free viewpoint video systems [10, 11], camera calibration is an indispensable step. Especially, for free viewpoint sports video, field model-based camera self-calibration is more flexible and popular. In automatic camera self-calibration, the homography matrix between a camera image and a field model is estimated from the camera image itself as shown in **Fig. 1**. Generally speaking, the conventional automatic self-calibration method involves two steps. The first step is to find crossing points of field lines in a camera image, and the second step estimates the homography matrix according to the correspondence of crossing points between a camera image and a field model. However, the main problem of such method is that exhaustive model matching has to be carried out to evaluate all the possible homography matrices and to select the best one. The processing speed of such an exhaustive calibration method is not sufficient for real-time application of a free viewpoint sports video system using a non-fixed camera. For instance, the conventional methods cannot synthesize free viewpoint video highlights of the first half of a soccer match during the half-time break.

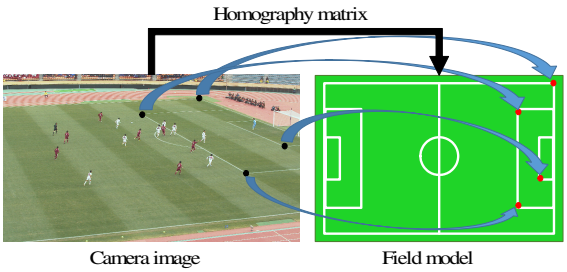


Fig. 1. Illustration of camera self-calibration for a soccer game. (The black points and red points are selected to estimate a homography matrix.)

In this paper, we consider a fast automatic camera self-calibration method for soccer videos based on a soccer field model, and we propose the construction of a new feature vector of crossing points of field lines in camera images and the field model. A large number of homography matrix estimations and evaluations can be thereby avoided according to the matching result of feature vectors. As far as we know, it is the first time to construct a feature vector for crossing points in camera calibration, and the proposed method reduces the number of estimations and evaluations of a homography matrix from hundreds of times to only several times.

2. RELATED WORK AND OUR CONTRIBUTIONS

There have been several studies on automatic camera self-calibration in sports video. Sudhir et al. [12] proposed a calibration method to detect several predefined points on a tennis court, but the method was not robust against the occlusions of the court lines. In addition, Alvarez [13] proposed a mathematical model for homography matrix estimation based on analyzing the center circle of a soccer field. However, the center circle was often not visible in camera images. A fundamental study was conducted by Farin et al. [14], where the proposed method was composed of white field line extraction and exhaustive model matching between the crossing points in a camera image and the predefined points in a field model to estimate all the possible homography matrices. However, the exhaustive model matching was very time consuming and impractical in real-time applications, especially when it involved the calibration of a non-fixed camera. Later, Farin et al. [15] proposed a fast calibration method which was based on RANSAC line parameter estimation. However, the RANSAC-based method still produced unreliable field lines, which led to inaccurate homography matrix. Recently, Yao [16] proposed an automatic camera self-calibration method that introducted a histogram of extracted field lines in a camera image to provide a robust calculation of crossing points and reduce the calibration time. However, similar to the work in [14] and [15], there was no additional information on the calculated crossing points, and hence exhaustive model matching was unavoidable.

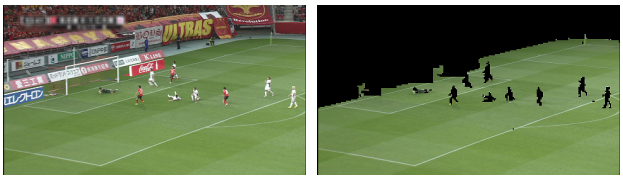
In this paper, based on our previous work [16], we propose the construction of a feature vector of crossing points in camera images and a field model. The directional pattern of a crossing point is analyzed and expressed as a numerical value. Therefore, a 4D feature vector is constructed by analyzing four crossing points in a camera image, and the 4D feature vector is matched with all the 4D feature vectors in the field model. According to the feature vector matching result, only a few homography matrices are estimated as the candidates and evaluated using a back projection method. Therefore, the number of homography matrix evaluations, which accounts for most of the calibration time, is greatly reduced. Experimental results show that the proposed method is effective and remarkably faster than other previous methods.

The remainder of this paper is organized as follows. In the third section, the proposed method is presented in detail. The experimental results are shown in section four, followed by a brief conclusion in the last section.

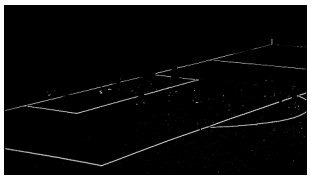
3. PROPOSED METHOD

3.1. Field Line Image Extraction

Basically speaking, as shown in **Fig. 2(a)**, the color of a soccer field in a camera image I^{RGB} is green and the color of the field lines is white. However, due to the lighting conditions,



(a) Camera image I (b) Field ground image G



(c) Field line image F

Fig. 2. Results of field line image extraction

color-based field extraction is too sensitive in an RGB color space. Thus, I^{RGB} is first converted to an HSV color space, written as $I^{RGB} \rightarrow I^{HSV}$. Next, several statistical thresholds, σ^H_{min}, σ^H_{max} and σ^V_{min} are estimated to produce a soccer field mask B by a labeling process

$$B(x,y) = \begin{cases} 1, I^H(x,y) \in [\sigma^H_{min}, \sigma^H_{max}] \&\& I^V(x,y) > \sigma^V_{min} \\ 0, \text{otherwise.} \end{cases} \quad (1)$$

where 1 indicates ground area while 0 indicates other area.

According to the labeled mask B, the soccer field ground image G is extracted, as shown in **Fig. 2(b)**. Following this, a field line image F is extracted on top of a field ground image G by white color detection [14], represented as

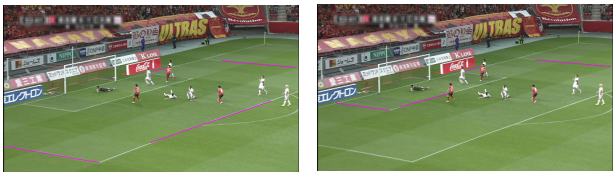
$$F(x,y) = \begin{cases} 255, & G(x,y) > th_w \&\& \Delta < th_{grad} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

, where the field ground image G has been converted into a gray image. th_w and th_{grad} are white color and pixel gradient thresholds, respectively. In addition, $\Delta = \mid G(x,y) - G(x+\tau,y) \mid + \mid G(x,y) - G(x-\tau,y) \mid + \mid G(x,y) - G(x,y+\tau) \mid + \mid G(x,y) - G(x,y-\tau) \mid$, and τ is the distance used to calculate the gradient of current pixel $G(x,y)$. Therefore, a clear field line image F is extracted and shown in **Fig. 2(c)**.

3.2. Feature Vector Generation and Matching

Our key idea is presented in this subsection, and the proposed feature vector construction of four crossing points is described below.

After a clear field line image is extracted, the Hough transform [17] is adopted to detect field lines in the field line image F. The histogram based line selection method in [16] is modified to obtain parameters (a slope and an offset) of field lines. In order to ensure that calculated crossing points are visible in the camera image for subsequent feature vector construction, more than two field lines are extracted first, and the field lines



(a) Field lines in [16] (b) Field lines in this study

Fig. 3. Field line extraction and selection (Selected field lines are highlighted in pink.)

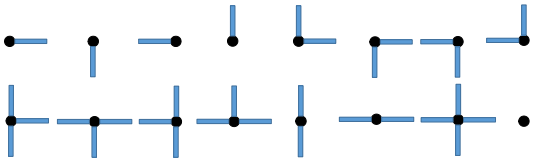


Fig. 4. The 16 possible patterns of a crossing point

that produce four visible crossing points are finally selected. A comparison between the selected field lines in [16] and in this study is shown in **Fig. 3**. It is clear that two crossing points of the selected field lines are invisible in **Fig. 3(a)**. In comparison, the selected field lines in this study produce four visible crossing points in the camera image.

It is observed that a crossing point of field lines can be distinguished by the information of the four directions as shown in **Fig. 4**. Therefore, a pattern of directional information for a crossing point is defined by a 4-bit value, where each bit value indicates whether one direction of the crossing point exists or not. Without losing generality, UP, RIGHT, DOWN, and LEFT are defined from the highest bit to the lowest bit. In addition, we assume that the camera shooting direction is known. Thus, the horizontal and vertical lines in a camera image are easily distinguished.

Based on the definition and assumption above, we compose a 4D feature vector to represent the four detected crossing points in a camera image as shown in **Fig. 5**. To be specific, the four crossing points (p^1, p^2, p^3, p^4) are formed as a point set $\mathbf{P} = \{p^1, p^2, p^3, p^4\}$ and sorted in a clockwise order according to the respective coordinates. In addition, the

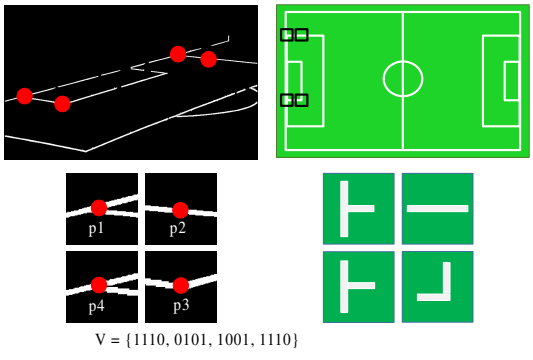


Fig. 5. The illustration of a pattern combination of crossing points in both camera image and field model

top-left point is always set as the starting one without losing generality. For each crossing point p^i, it is acknowledged that the information of two field lines l_v^i, l_h^i is available, and thus we check the four directions of the crossing point p^i along the directions of l_v^i and l_h^i. According to the pattern definition mentioned above, the extracted patterns of four crossing points in **Fig. 5** are characterized as a 4D feature vector $\mathbf{V} = \{1110, 0101, 1001, 1110\}$ in binary format or written as $\mathbf{V} = \{14, 5, 9, 14\}$ in decimal format. Similarly, various 4D feature vectors of points in a field model can be generated and prepared in advance.

Therefore, the exhaustive homography evaluation is simplified as feature vector matching, which can greatly reduce the complexity. Feature vector matching is represented as

$$\mathbf{V}_{\text{cand}} = \mathbf{V}_t, \quad \text{if } \| \mathbf{V} - \mathbf{V}_t \|_{l_2} < \epsilon, \tag{3}$$

where $\mathbf{V}$ is the feature vector of the 4-point set $\mathbf{P}$ in a camera image and $\mathbf{V_t}$ is the feature vector of a 4-point set $\mathbf{P_t}$ in a field model. In addition, T is the number of selected $\mathbf{V}_{\text{cand}}$ that keeps the matching error within a threshold ϵ, where $\epsilon = 0$ if an exact matching is required and more $\mathbf{V}_{\text{cand}}$ will be selected if ϵ is relaxed to a positive value. According to the result of feature vector matching, $\mathbf{P}$ and the selected $\mathbf{P}_{\text{cand}}$ are adopted in homography matrix estimation.

3.3. Homography Matrix Estimation and Evaluation

It is noted that a homography matrix, H, projects a point p in a plane to a point q in another plane [18]. Therefore, we assume $p(x, y, z)$ as a crossing point in a camera image and correspondingly $q(X, Y, Z)$ as a point in a field model. Thus, there is

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \sim \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{4}$$

If we define a normalization as $x' = x/z$, $y' = y/z$ and $X' = X/Z$, $Y' = Y/Z$, there is

$$\begin{aligned} x' &= \frac{x}{z} = \frac{h_{11}X' + h_{12}Y' + h_{13}}{h_{31}X' + h_{32}Y' + h_{33}} \\ y' &= \frac{y}{z} = \frac{h_{21}X' + h_{22}Y' + h_{23}}{h_{31}X' + h_{32}Y' + h_{33}} \end{aligned} \tag{5}$$

Therefore, considering a homography matrix with eight degrees of freedom, four pairs of point correspondences between $\mathbf{P}$ and $\mathbf{P}_{\text{cand}}$ are adopted to calculate H.

Moreover, according to **eq.(3)**, there might be multiple candidates of feature vectors $\mathbf{V}_{\text{cand}}$ and point sets $\mathbf{P}_{\text{cand}}$. Therefore, an evaluation method is adopted to finally obtain the optimal homography matrix. It is assumed that the points in field lines of a field model should be projected onto the white lines of the corresponding field line image if there exists an accurate homography matrix. Therefore, a penalty value

d_k of each projected point $(\hat{x}, \hat{y})$ in the field line image F is calculated as

$$\begin{cases} d_k = 1, & \text{if } F(\hat{x}, \hat{y}) \geq th_p \\ d_k = -2, & \text{otherwise,} \end{cases} \tag{6}$$

where th_p is the threshold value. Based on the definition of d_k in **eq.(6)**, the optimal homography matrix H_{opt} yields to

$$H_{opt} = \arg\max_{H} \sum_{1 \leq k \leq K} d_k, \tag{7}$$

where K is the total number of points in the field model.

4. EXPERIMENTAL RESULTS

In order to verify the effectiveness of our proposed method, we carried out tests on two soccer scenes captured by PTZ cameras. The resolution of the camera is full HD and the frame rate is 30fps. The parameters are set as follows. In the field line image extraction, σ^H_{min}, σ^H_{max} and σ^V_{min} are estimated as 30, 75 and 40 respectively. In addition, two thresholds th_w and th_{grad} are set as 145 and 40, and the value τ is 4 in pixel gradient calculation. Finally, the threshold th_p is set as 127 in **eq.(6)**. All the parameters are fixed throughout the entire experiment.

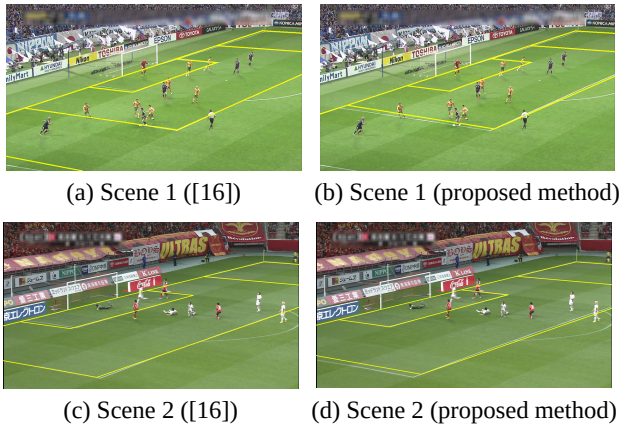
Th other methods in [14] and [16] are compared with our proposed method. The camera shooting direction is prior information for the three methods. Basically, homography matrix evaluation accounts for most of the time in camera self-calibration. Therefore, for simplicity, we consider the number of homography matrix evaluations as the cost of calibration.

The comparison of calibration cost is shown in **Table 1**. The second row in **Table 1** shows the number D of feature vectors in a camera image. The method in [16] and our proposed method detected two horizontal and two vertical lines in the calculation of crossing points in a camera image. Thus, there is only one feature vector in camera image. In comparison, the method in [14] detected at least three vertical and three horizontal lines, producing at least nine possible feature vectors. The third row in **Table 1** shows the number, T of feature vector candidates in a soccer field model. For an exhaustive search, $T = \binom{6}{2} \times \binom{7}{2} = 315$ because there are 6 horizontal lines and 7 vertical lines in the soccer field model as shown in **Fig. 1**. In comparison, an exhaustive search is not necessary in our proposed method, and the specific number of searches depends on the 4D feature vector of a camera image. According to the value of $\epsilon = 10$ in our experiment, T may be different numbers, such as 4, 9, 11, and 16. Therefore, $T \leq \mathbf{20}$ is a reasonable result for the proposed method. Furthermore, the bottom row in **Table 1** shows the running time (ms) required to calibrate one frame (including homography matrix estimation and evaluation), and it is verified that the proposed method is much faster than the other two methods.

Finally, field model projection is adopted to evaluate the calibration accuracy. The projected field lines are marked in yellow, and the comparison between the method described in [16] and the proposed method is shown in **Fig. 6**. Basically, the accuracy of the proposed method is slightly lower than the accuracy in [16], and the main reason is that the selected crossing points in the proposed method are limited in a small area (goal area). However, such slight loss of calibration accuracy has virtually no effect on the final synthesized free viewpoint soccer video, and note that calibration accuracy will be increased if the camera is zoomed out because crossing points in a large area will be visible and used in the proposed calibration method.

Table 1. A comparison of the calibration cost between the proposed method and other methods for one frame

Calibration cost	[14]	[16]	Proposal
# of $\mathbf{V}$ in image (D)	9	1	**1**
# of $\mathbf{V}_{\text{cand}}$ (T)	315	315	$\leq$ **20**
# of H evaluation	315×9	315	$\leq$ **20**
Running time	18326 ms	1940 ms	**85** ms



(a) Scene 1 ([16]) (b) Scene 1 (proposed method)

(c) Scene 2 ([16]) (d) Scene 2 (proposed method)

Fig. 6. Comparison of calibration accuracy between of the proposed method and previous method

5. CONCLUSION

In this paper, a fast self-calibration method for a non-fixed camera was proposed in order to synthesis free viewpoint soccer videos. Different from any previous methods, as far as we know, it was the first time to construct a feature vector of crossing points, and a 4D feature vector was constructed by analyzing the directional information of four crossing points. Therefore, based on the result of feature vector matching, it was not necessary to exhaustively calculate and evaluate all the possible homography matrices, and hence the calibration cost was greatly reduced. The experimental results showed that the proposed method was much faster than other methods with a slight loss of calibration accuracy that was negligible in the final synthesized free viewpoint soccer video.

6. REFERENCES

[1] Tanimoto M., Panahpour Tehrani M., Fujii T., Yendo T., "Free-viewpoint TV," *Signal Processing Magazine, IEEE*, vol. 28, no.1, pp. 67–76, 2011.

[2] Carranza J., Theobalt C., Magnor M.-A., and Seidel H.-P, "Free-viewpoint video of human actors," *In ACM transactions on graphics (TOG)*, Vol. 22, No. 3, pp. 569–577, ACM., 2003, July.

[3] Smolic A., Mueller K., Merkle P., Fehn C., Kauff P., Eisert P., and Wiegand T. "3D video and free viewpoint video-technologies, applications and MPEG standards," *In 2006 IEEE International Conference on Multimedia and Expo*, pp. 2161–2164, 2006, July.

[4] Kilner, J.-J., Starck J.-R. , and A. Hilton, "A Comparative Study of Free Viewpoint Video Techniques for Sports Events," *IET European Conference on Visual Media Production. IET*, 2006.

[5] Suenaga R., Suzuki K., Tezuka T., Panahpour Tehrani M., Takahashi K., and Fujii T., "A practical implementation of free viewpoint video system for soccer games," *IS& T/SPIE Electronic Imaging International Society for Optics and Photonics*, 2015.

[6] Yao Q., Sankoh H., Sabirin H., and Naito S., "Accurate silhouette extraction of multiple moving objects for free viewpoint sports video synthesis," *Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on*, 2015.

[7] Yao Q., Nonaka K., Sankoh H., and Naito S., "Robust moving camera calibration for synthesizing free viewpoint soccer video," *In 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1185–1189, 2016, September.

[8] Ohta Y., et al., "Live 3D video in soccer stadium," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 173–187, 2007.

[9] Bal B., and Dureja G., "Hawk eye: a logical innovative technology use in sports for effective decision making," *Sport Science Review*, Vol. 21 No.1-2, pp. 107–119, 2012.

[10] Decoret X., et al., "Billboard clouds for extreme model simplification," *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3, pp. 689–696, 2003.

[11] Grau, Oliver, et al., "A free-viewpoint video system for visualization of sport scenes," *SMPTE motion imaging journal*, Vol. 116, No. 5-6 pp. 213–219, 2007.

[12] Sudhir G., et al., "Automatic classification of tennis video for high-level content-based retrieval," *Content-Based Access of Image and Video Database, 1998 IEEE International Workshop on*, pp. 81–90, 1998.

[13] Alvarez L., and Caselles V. "Homography estimation using one ellipse correspondence and minimal additional information," *In 2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4842–4846, 2014.

[14] Farin D., Susanne K., and Wolfgang E., "Robust camera calibration for sport videos using court models," *SPIE Electronic Imaging 2004. International Society for Optics and Photonics*, pp. 80–91, 2004.

[15] Farin D., Han J., Peter H.N. DE WITH, "Fast camera calibration for the analysis of sport sequences," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, IEEE, 2005.

[16] Yao Q., Sankoh H., Nonaka K., and Naito S., "Automatic Camera Self-Calibration for Immersive Navigation of Free Viewpoint Sports Video," *Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on*, 2016.

[17] Kiryati N., Eldar Y., and Bruckstein A.-M., "A probabilistic Hough transform," *Pattern recognition*, Vol. 24, No.4, pp. 303–316, 1991.

[18] Zhang Z., "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 22, No. 11, pp. 1330–1334, 2000.